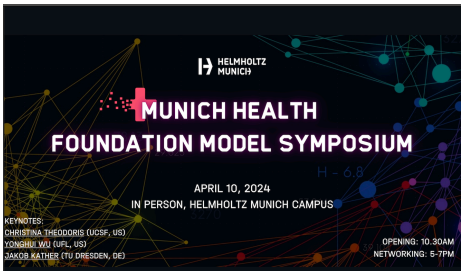

Contribution ID: **8** Type: **Talk**

Insights from Acquiring Open Medical Imaging Datasets for Foundation Model Development

Foundation Models require significantly more data for training than earlier AI generations. The scarcity of clinical data, as well as the necessity of perfecting model generalization capabilities, make it necessary to aggregate data for model training and validation from various datasets. In this work, we explore the challenges related to FAIR clinical imaging data (findability, accessibility, interoperability, and reusability) encountered while sourcing real open clinical imaging datasets from large public cohort studies, existing public data repositories and individual dataset publications. Additionally, we present anonymized real-world examples detailing access, metadata, and licensing configurations, illustrating specific problems that may emerge with regard to various FAIR principles. We introduce a tier system designed to identify dataset issues impacting machine readability. Furthermore, we evaluate the manual efforts and resources required to find, access, and fetch data for Foundation Model training, linking these activities to our tier-based framework for assessing dataset machine readability. Additionally, we provide some suggestions on how to refine datasets on different tiers to make compliant with the FAIR criteria and hence reduce the human workload of their procurement. Key strategies, such as utilizing Resource Description Framework to export key-value pairs from the Imaging Data Repository and constructing FAIR Data Point, are given as methods to facilitate highly automated dataset access through advanced techniques that adhere to FAIR standards.

Primary authors: Mr ULRICH, Constantin (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany); Dr ISENSEE, Fabian (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany); Mr MOORE, Josh (German BioImaging, University of Konstanz, Germany); Prof. MAIER-HAIN, Klaus (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany); Mr KULLA, Lucas (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Helmholtz Metadata Collaboration (HMC) Hub Health, German Cancer Research Center (DKFZ), Heidelberg, Germany); Dr NOLDEN, Marco (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany; Helmholtz Metadata Collaboration (HMC) Hub Health, German Cancer Research Center (DKFZ), Heidelberg, Germany); Dr JÄGER, Paul (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany); Mr SCHADER, Philipp (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany; Helmholtz Metadata Collaboration (HMC) Hub Health, German Cancer Research Center (DKFZ), Heidelberg, Germany); Mr DVORETSKII, Stefan (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg Ger-

man Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Helmholtz Metadata Collaboration (HMC) Hub Health, German Cancer Research Center (DKFZ), Heidelberg, Germany); Mr WALD, Tassilo (Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany)

Session Classification: Break + Posters session

Track Classification: Poster